UMUX-LITE – When There's No Time for the SUS

James R. Lewis
IBM Software Group
Boca Raton, FL
jimlewis@us.ibm.com

Brian S. Utesch
IBM Software Group
Durham, NC
butesch@us.ibm.com

Deborah E. Maher
IBM Software Group
Cambridge, MA
debmaher@us.ibm.com

ABSTRACT
In this paper we present the UMUX-LITE, a two-item questionnaire based on the Usability Metric for User Experience (UMUX) [6]. The UMUX-LITE items are "This system's capabilities meet my requirements" and "This system is easy to use." Data from two independent surveys demonstrated adequate psychometric quality of the questionnaire. Estimates of reliability were .82 and .83 – excellent for a two-item instrument. Concurrent validity was also high, with significant correlation with the SUS (.81, .81) and with likelihood-to-recommend (LTR) scores (.74, .73). The scores were sensitive to respondents' frequency-of-use. UMUX-LITE score means were slightly lower than those for the SUS, but easily adjusted using linear regression to match the SUS scores. Due to its parsimony (two items), reliability, validity, structural basis (usefulness and usability) and, after applying the corrective regression formula, its correspondence to SUS scores, the UMUX-LITE appears to be a promising alternative to the SUS when it is not desirable to use a 10-item instrument.

Author Keywords
System Usability Scale; SUS; Usability Metric for User Experience; UMUX; UMUX-LITE; psychometric evaluation; usability evaluation; standardized questionnaires; satisfaction measures

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces–Evaluation/Methodology.

General Terms
Human Factors; Design; Measurement.

INTRODUCTION

Research Motivation
A typical summative usability test includes the assessment of satisfaction along with assessments of effectiveness and efficiency [7, 14]. Starting in the late 1980s, standardized usability questionnaires appropriate for usability testing began to appear [14]. One of the most popular of these is the System Usability Scale (SUS), accounting for an estimated 43% of post-test questionnaire usage [12].

With just ten items, the SUS is fairly short, but in our practice we have encountered situations in which an even more concise questionnaire is desirable. This is especially the case when post-test debriefing involves a large number of questions or when the satisfaction questionnaire is part of a much larger survey. For this reason we were intrigued when we came across the Usability Metric for User Experience (UMUX) [6] – a four-item questionnaire claimed to be an effective proxy for the SUS.

Despite the initial research supporting the use of the UMUX as a proxy for the SUS [6], a recent review of the UMUX [8] raised several criticisms, including:

- How much time do respondents really save when answering four rather than ten questions?
- A parallel analysis [4] of the eigenvalues from a principal components analysis of UMUX scores suggested a bidimensional rather than the claimed unidimensional structure.

Because we routinely use the SUS, we decided to continue collecting SUS scores while simultaneously collecting UMUX scores in pursuit of two research goals:

1. Attempt to replicate the results reported in the original UMUX research [6].
2. Conduct additional item and structural analyses to investigate the feasibility of further reducing the number of UMUX items to use in a quickly-conducted unidimensional surrogate of the SUS – a UMUX-LITE.

Next, we provide summaries of the key properties and prior psychometric research of the SUS and UMUX, followed by analyses and conclusions based on new data from two independent surveys.

The System Usability Scale (SUS)
The SUS is a ten-item questionnaire using five-point scales. Responses to SUS items are recoded to produce an overall SUS score that ranges from 0 to 100 in 2.5 point increments. Although a self-described "quick-and-dirty" questionnaire [3], the SUS appears to have excellent psychometric properties (estimates of reliability typically exceeding 0.9, significant concurrent validity with ratings

CHI 2013, April 27–May 2, 2013, Paris, France.

of user friendliness, and sensitivity to variables such as system and frequency of use) [1, 2, 9].

The SUS is available in two versions. The Standard version has items with mixed tone – odd items have a positive tone; even items have a negative tone. In the Positive version, all items have a positive tone. Sauro and Lewis [13] found that the Positive version had advantages over the Standard version with regard to reductions in misinterpretation, mistakes, and miscoding. Both versions had high reliability (Standard: 0.92; Positive: 0.96), and had no significant difference in their mean scores. There was no evidence of acquiescence or extreme response biases in the Positive version. Table 1 shows the item content for both versions of the SUS.

Although generally treated as a unidimensional measure, recent analyses suggest that the SUS is more likely a bidimensional measure, with factors associated with the constructs of Usable (Items 1-3, 5-9) and Learnable (Items 4, 10) [2, 9].

Item	Standard	Positive
1	I think that I would like to use this system frequently.	I think that I would like to use this system frequently.
2	I found the system unnecessarily complex.	I found the system to be simple.
3	I thought the system was easy to use.	I thought the system was easy to use.
4	I think that I would need the support of a technical person to be able to use this system.	I think I could use the system without the support of a technical person.
5	I found the various functions in the system were well integrated.	I found the various functions in the system were well integrated.
6	I thought there was too much inconsistency in this system.	I thought there was a lot of consistency in the system.
7	I would imagine that most people would learn to use this system very quickly.	I would imagine that most people would learn to use this system very quickly.
8	I found the system very cumbersome to use.	I found the system very intuitive.
9	I felt very confident using the system.	I felt very confident using the system.
10	I needed to learn a lot of things before I could get going with this system.	I could use the system without having to learn anything new.

Table 1. Standard and Positive Versions of the SUS.

The Usability Metric for User Experience (UMUX)

The UMUX [6] is a relatively new standardized usability questionnaire designed to get a measurement of perceived usability consistent with the SUS, but using fewer items that more closely conformed to the ISO definition of usability (effective, efficient, satisfying) [7]. UMUX items vary in tone and have seven scale steps from 1 (strongly disagree) to 7 (strongly agree). Starting with an initial pool of 12 items, the final UMUX had four items that included a general question similar to the Single Ease Question ("This system is easy to use") [11] and the best candidate item from each of the item sets associated with efficiency, effectiveness, and satisfaction, where "best" meant the item with the highest correlation to the concurrently collected overall SUS score. Using a recoding scheme similar to the SUS, a UMUX score can range from 0 to 100. The four UMUX items are:

1. This system's capabilities meet my requirements.
2. Using this system is a frustrating experience.
3. This system is easy to use.
4. I have to spend too much time correcting things with this system.

To validate the UMUX, Finstad (its developer) [6] had users of two systems, one with a reputation for poor usability (System 1, n = 273) and the other perceived as having good usability (System 2, n = 285), complete the UMUX and the Standard SUS. As expected, the reliability of the SUS was high, with a coefficient alpha of 0.97. The reliability of the UMUX was also high, with a coefficient alpha of 0.94. The UMUX scores for the two systems were significantly different ($t(533) = 39.04$, $p < 0.01$) with System 2 getting better scores than System 1 (evidence of sensitivity). More importantly, there was an extremely high correlation between the SUS and UMUX scores ($r = 0.96$, $p < 0.001$), providing evidence of strong concurrent validity and suggesting that the UMUX was statistically equivalent to the SUS.

METHOD

As part of two independent surveys, we had an opportunity to simultaneously capture responses to the SUS, the UMUX, and a likelihood-to-recommend (LTR) item. Respondents were IBM employees with varying amounts of experience with the evaluated system (from using the system once every few months to more than once a day). In one survey, respondents completed the Positive version of the SUS (n = 402); in the other they completed the Standard version (n = 389).

RESULTS

UMUX Item Analysis

Table 2 shows the correlations (including 99% confidence intervals) for the UMUX items with the Positive and Standard versions of the SUS. Across the datasets the intervals were identical for the odd-numbered (positive tone) items. The correlations with the negative tone items, in contrast, were significantly different between the positive and standard versions of the SUS. Item 4 had the lowest correlation with the SUS in both datasets.

Item	r(Standard)	r(Positive)
1	.69 - **.75** - .81	.69 - **.75** - .80
2	.75 - **.80** - .84	.52 - **.60** - .68
3	.76 - **.81** - .85	.76 - **.81** - .85
4	.57 - **.66** - .72	.27 - **.38** - .49

Table 2. UMUX Item to SUS Correlations (with 99% confidence intervals).

UMUX and SUS Factor Analyses

Table 3 shows the results of a factor analysis of the UMUX items combined across the datasets (analyses by dataset showed the same pattern). As predicted in the review of the original UMUX research [8], the UMUX had a clear bidimensional structure with positive-tone items aligning with one factor and negative-tone items aligning with the other, a solution supported by parallel analysis of the eigenvalues [4].

Item	Tone	Factor 1	Factor 2
1	Pos	**0.762**	0.295
2	Neg	0.485	**0.716**
3	Pos	**0.776**	0.393
4	Neg	0.235	**0.659**

Table 3. Factor Analysis of the UMUX.

Item	Factor 1	Factor 2
1	0.351	0.203
2	**0.731**	0.326
3	**0.697**	0.341
4	0.226	**0.698**
5	**0.734**	0.209
6	**0.668**	0.252
7	**0.653**	0.403
8	**0.743**	0.412
9	**0.600**	0.507
10	0.330	**0.634**

Table 4. Factor Analysis of the SUS.

Table 4 shows the results of a factor analysis of the SUS items combined across the datasets (analyses by dataset showed the same pattern). The results essentially replicated the results of previous analyses showing a two-factor structure [2, 9], with one exception -- Item 1 didn't strongly associate with either factor (but did have a higher loading on the first factor, consistent with earlier findings).

Psychometric Quality of the UMUX

In general, our results replicated the findings reported by Finstad [6]. For the two datasets, the UMUX correlated significantly with the SUS (with 99% confidence intervals; Standard: .87 - **.90** - .92; Positive: .74 - **.79** - .84). Although this is significantly less than the originally claimed correlation of .96 (99% confidence interval ranging from 0.95 to 0.97), it is evidence of concurrent validity. The estimated reliabilities of the UMUX were adequate (.87, .81), but like the correlations with the SUS, quite a bit less than the originally reported value of .97. For both datasets, there was no significant difference between the mean SUS and mean UMUX scores (extensive overlap between the 99% confidence intervals), consistent with the original data.

Potential UMUX Variants

There are many potential variants of the UMUX, but based on the item and factor analyses above, two stand out. One variant would be to drop Item 4 due to its relatively low correlation with the SUS, leaving three items. A second would be to drop the negative-tone items leaving the two positive-tone items, Item 1 associated with usefulness (functional adequacy) and Item 3 associated with usability (ease-of-use). Despite the common wisdom that attitudinal questionnaires should contain a mix of positively- and negatively-toned items, there is a body of research that argues against this practice [5, 10, 13, 15, 16].

We decided to pursue the second variant composed of the positive-tone UMUX items. The primary reasons for this choice were the parsimony of the resulting instrument (two items) and its connection through the content of the items to the Technology Acceptance Model [5], a questionnaire from the market research literature that assesses the usefulness and ease-of-use of systems, and has an established relationship to likelihood of future use.

Psychometric Assessment of the UMUX-LITE

Psychometric analyses using just the positive-tone items of the UMUX indicated good psychometric quality. Like the full UMUX, this metric correlated significantly with both Standard and Positive versions of the SUS (.81, .85) and with LTR (.73, .74), evidence of concurrent validity. Coefficient alpha indicated acceptable scale reliability (> .7) for both datasets (.83, .82).

Unlike the full UMUX scores, there was a small but statistically significant difference between the overall SUS scores and the scores based just on UMUX Items 1 and 3. To compensate for that difference, we used linear regression on the combined samples (n = 791) to compute the UMUX-LITE:

$$\text{UMUX-LITE} = .65(\text{UMUX}_{(1,3)}) + 22.9$$

In that formula, ($\text{UMUX}_{(1,3)}$) refers to a UMUX score computed from just Items 1 and 3, using a SUS-like procedure to obtain a score that ranges from 0 to 100 (specifically, subtract 1 from each 7-point item, add them together, then multiply by 100/12). Applying the regression equation to compute the UMUX-LITE from $\text{UMUX}_{(1,3)}$ brought the UMUX-LITE scores into correspondence with the SUS scores for both datasets.

DISCUSSION

We set out to see if an independent assessment of the UMUX would lead to the same results as the original research that produced the UMUX [6]. We replicated many of the original findings with regard to typical goals of psychometric evaluation (adequate reliability and validity), although our estimates of reliability and validity tended to be lower than those from the original research. One notable exception was that our structural analysis indicated that the UMUX is bidimensional rather than unidimensional, with items aligning on factors as a function of the tone of the item (positive/negative).

We also wanted to see if an even shorter questionnaire based on the UMUX would have acceptable psychometric properties and, using item and structural analysis, settled on an instrument based on its positive-tone items – the UMUX-LITE. This two-item instrument (with adjustment based on a regression equation to match it to the SUS) had acceptable reliability and validity – in fact, its psychometric properties were very good given it only has two items.

LIMITS TO GENERALIZABILITY AND FUTURE WORK

Although these results are encouraging, it is important to keep in mind that this is just a first step. In contrast to the relatively rich research literature on the SUS, the only published research to date for the UMUX is the original paper [6] and this one, and this paper is the first to provide psychometric properties of the UMUX-LITE. Until researchers have validated the UMUX-LITE across a wider variety of systems, we do not recommend its use independent of the SUS. Given the promising results so far, however, we do recommend that practitioners and researchers who use the SUS include the UMUX-LITE items in their work to begin building independent databases for future evaluation of its reliability, validity, and sensitivity. We certainly intend to do so.

REFERENCES

1. Bangor, A., Kortum, P.T., and Miller, J.T. An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, 24 (2008), 574-594.
2. Borsci, S., Federici, S., and Lauriola, M. On the dimensionality of the system usability scale: A test of alternative measurement models. *Cognitive Processes*, 10 (2009), 193-197.
3. Brooke, J. SUS: A "quick and dirty" usability scale. In: Jordan, P., Thomas, B., Weerdmeester, B. (Eds.), *Usability Evaluation in Industry*. Taylor & Francis, London, UK, (1996) 189–194.
4. Coovert, M.D., and McNelis, K. Determining the number of common factors in factor analysis: A review and program. *Educational and Psychological Measurement*, 48 (1988), 687-693.
5. Davis, D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13 (1989), 319-339.
6. Finstad, K. The usability metric for user experience. *Interacting with Computers*, 22 (2010), 323-327.
7. ISO 9241-11. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). Part 11: Guidance on Usability, 1998.
8. Lewis, J. R. Critical review of "The Usability Metric for User Experience". *Interacting with Computers* (In press).
9. Lewis, J.R., and Sauro, J. The factor structure of the System Usability Scale. In: Kurosu, M. (Ed.), *Human Centered Design, HCII 2009*. Springer-Verlag, Heidelberg, Germany, (2009) 94-103.
10. Pilotte, W.J., and Gable, R.K. The impact of positive and negative item stems on the validity of a computer anxiety scale. *Educational and Psychological Measurement*, 50 (1990), 603-610.
11. Sauro, J., and Dumas, J.S. Comparison of three one-question, post-task usability questionnaires. In *Proc. CHI 2009*, ACM Press (2009), 1599-1608.
12. Sauro, J., and Lewis, J.R. Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proc. CHI 2009*, ACM Press (2009), 1609-1618.
13. Sauro, J., and Lewis, J.R. When designing usability questionnaires, does it hurt to be positive? In *Proc. CHI 2011*, ACM Press (2011), 2215-2223.
14. Sauro, J., and Lewis, J. R. *Quantifying the User Experience: Practical Statistics for User Research*. Morgan-Kaufmann, Waltham, MA, USA, 2012.
15. Schmitt N., and Stuits, D. Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 9 (1985), 367-373.
16. Stewart, T.J., and Frye, A.W. Investigating the use of negatively-phrased survey items in medical education settings: Common wisdom or common mistake? *Academic Medicine*, 79 (2004), S1-S3.